Post-Print (final draft post-refereeing)
Published article: Cole, T.W. & Shreeves, S.L. (2004). Search and discovery across collections: The IMLS Digital Collections and Content Project. Library Hi Tech 22(3): 307-322 DOI: 10.1108/07378830410560107.

Search and Discovery across Collections: the IMLS Digital Collections and Content Project

AUTHORS

Timothy W. Cole is Mathematics Librarian and Professor of Library Administration at the University of Illinois at Urbana-Champaign where he has been a member of the Library faculty since 1989. He has held prior appointments at Illinois as Assistant Engineering Librarian for Information Services and Systems Librarian for Digital Projects. He is Principal Investigator for the University of Illinois IMLS Digital Collections and Content Project.

E-mail: t-cole3@uiuc.edu

Sarah L. Shreeves is Visiting Assistant Professor of Library Administration and Project Coordinator for the University of Illinois IMLS Digital Collections and Content Project. Previously she was a Project Coordinator for the University of Illinois Open Archives Initiative Metadata Harvesting Project funded by the Andrew W. Mellon Foundation. From 1992 until 2001 she was a member of staff at the Massachusetts Institute of Technology Libraries.

Email: sshreeve@uiuc.edu

Abstract

In the fall of 2002 the University of Illinois Library at Urbana-Champaign received a grant from the Institute of Museum and Library Services (IMLS) to implement a collection registry and item-level metadata repository for digital collections and content created by or associated with projects funded under the IMLS National Leadership Grant (NLG) program. When built, the registry and metadata repository will facilitate retrieval of information about digital content related to past and present NLG projects. The process of creating these services also is allowing us to research and gain insight into the many issues associated with implementing such services and the magnitude of the potential benefit and utility of such services as a way to connect, bring together, and make more visible a broad range of heterogeneous digital content. This paper describes the genesis of our project, the rationale for architectural design decisions, challenges faced, and our progress to date.

Keywords

Collection-level description; Collection registries; Metadata aggregations and repositories; Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH); use of Dublin Core.

Introduction

The World Wide Web offers cultural heritage institutions opportunities to enhance end-user services and reach larger and more widely distributed constituencies. Over the past few years there has been an explosion in the number of online information resources implemented by museums, libraries, archives, historical societies, and other cultural heritage institutions as they attempt to more aggressively exploit the potential of the Web. The benefit of having a rich diversity of quality and authoritative information available online is clear, but the magnitude of that benefit is tempered for many end-users by the difficulties in locating specific, desired information resources within the almost overwhelming aggregation of information now available. Every week there is more useful information available to find, but also every week, the amount of information that must be sorted through to find specific information desired grows as well. (Lyman & Varian, 2003) In addition, much of the information is 'hidden' or 'invisible', i.e. in databases and other locations less accessible to Web search engines. (Sherman & Price, 2003) The community continues to struggle to develop new techniques for managing the glut of information and to transform traditional methods of curation and librarianship in order to better organize available digital information in aggregate and make it easier for end-users to find the specific online information they want and need to answer specific questions.

In 2001 the Institute of Museum and Library Services (IMLS) [1] commissioned a Digital Library Forum to "discuss the implementation and management of networked digital libraries, including issues of infrastructure, metadata, thesauri and other vocabularies, and content enrichment such as curriculum materials and teacher guides." (IMLS 2001) In particular, the IMLS asked Forum members to examine and comment on opportunities for bringing the rich collections created with IMLS funding into digital libraries of national scope, an exemplar of which was (and is) the National Science Foundation's National Science Digital Library (NSDL) [2]. The report of the Forum, developed with significant input from several NSDL participants, included general recommendations to the IMLS as well as specific recommendations for projects funded by IMLS. The IMLS Forum also developed and promulgated a *Framework of Guidance for Building Good Digital Collections* (Cole, 2002), which has since been adopted by the National Information

Standards Organization (NISO). [3] Among the general suggestions to IMLS, the Forum recommended that IMLS "should maintain its own registry of funded digital collections." (IMLS 2001) Acting on this recommendation and on other input, the IMLS in the fall of 2002 funded the University of Illinois Library at Urbana-Champaign to research, design, develop, and demonstrate a pilot implementation of a collection registry and item-level metadata repository based on the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [4] to hold information describing digital collections and content created by or directly associated with National Leadership Grant (NLG) projects funded by IMLS since the inception of the NLG program in 1998.

The Illinois research and demonstration project, now at its midpoint, is the subject of this paper. We describe in order the motivations for and objectives of our project, specific high-level architecture design decisions made, the nature of the challenges we have encountered, and our accomplishments to date. Along the way we discuss lessons learned to date and consider the relevance of this project to similar work going on elsewhere in the U.S. and in Europe. We conclude with a brief discussion of open issues and planned work through the rest of our project.

Project Rationale, Goals, & Objectives

In undertaking to provide a collection registry and item-level metadata repository for digital collections and content associated with IMLS NLG projects, there are two levels of motivation. On the one hand, a NLG registry and metadata repository can serve several immediate needs and parochial interests of the IMLS. In its final report the IMLS Digital Library Forum suggested that a registry of IMLS funded digital content "can aid grant applicants looking for models and practical examples of acceptable practice, can help further the sense of community among past and present awardees, and can provide a mechanism for identifying collections with various features (for example, those existing collections which might be appropriate for future inclusion in the NSDL)." (IMLS, 2001) To that we would also add that a NLG registry and metadata repository can provide a more comprehensive view of some of the best products and outcomes of the IMLS NLG program, which in turn would be useful for those in the field and in the general public looking for a single entry point through which they can learn more about the scope and accomplishments of IMLS funded digitization and content management programs.

But our project is also a NLG project in its own right, and so is designed to demonstrate more generally certain infrastructure components potentially useful for building dispersed and dynamic user-centric digital library services. In this regard, our project is informed by and seeks to test in practice certain assumptions and hypotheses about the organization of digital content and the way in which such content can be shared and accessed effectively. The *Framework of Guidance for Building Good Digital Collections* draws an implicit distinction between digital collections and the value-added services which define digital libraries. The *Framework* articulates in one of its principles for "good" digital collections the definition that, "a good [digital] collection fits into the larger context of significant related national and international digital library initiatives." The *Framework* expands on this statement of principle in its introduction:

> In today's digital environment, the context of content is a vast international network of digital materials and services. Objects, metadata and collections should be viewed not only within the context of the projects that created them but as building blocks that others can reuse, repackage, and build services upon. Indicators of goodness correspondingly must now also emphasize factors contributing to interoperability, reusability, persistence, verification and documentation.
>
> (NISO, 2004)

This view of information collections and objects as reusable building blocks or "recombinant" components of broader information systems (Seaman, 2003), and the distinction between digital collections and information objects and the value-added services that access and make use of them (Dempsey, 2003; Lynch, 2002), is consistent with a model of collections suggested by earlier researchers which describes bodies of information content as "information landscapes." As defined by Michael Heaney,

> The information landscape can be seen as a contour map in which there are mountains, hillocks, valleys, plains and plateaux…. A specialized collection of particular importance is like a sharp peak. Upon a plateau there might be undulations representing strengths and weaknesses…. The landscape is, however, multidimensional. Where one scholar may see a peak another may see a trough. The task is to devise mapping conventions which enable scholars to read the map of the landscape fruitfully, at the appropriate

> level of generality or specificity. (Heaney, 2000)

If these views of scholarly collections and content are correct, an essential role of digital libraries must be to offer the value-added services that provide the dynamic mapping functions described by Heaney, so as to allow scholars and students to view the information landscapes they encounter in the most useful ways possible to each individually. We hypothesize that infrastructure components such as collection registries and item-level metadata aggregations, assuming they are populated with collection-level and item-level descriptive metadata records of adequate quality, can support these essential classes of digital library services. It is our view that such infrastructure components have the potential to facilitate the reuse of digital content in new and different ways – by enabling more effective search and discovery across multiple collections, and by supporting the kinds of dynamic mapping between collections and among and between individual information objects that will allow communities of scholarly interest to view an information landscape as best meets their needs. Thus a second rationale for our project is to create a testbed suitable for examining this hypothesis and the degree to which it might be valid within selected communities of interest.

As discussed below, other projects as well are looking at similar models and hypotheses, but not all researchers in the community favor collection registries and heterogeneous metadata aggregations as cost effective ways to map the information landscape. Standardized approaches to collection-level description in particular have not been well explored or tested in the United States as a piece of the cross-domain resource discovery puzzle. At one end of the spectrum of opinion there are concerns that approaches relying on ad-hoc connections between relatively small dispersed collections and dynamic recombinant approaches for associating widely dispersed and heterogeneous information objects lack the scale, sophistication, and tight coupling to specific target audiences necessary to sustain digital collections and make them truly useful in a scholarly context. In *First Monday*, Donald Waters of the Mellon Foundation, summarizing in narrative form the main points of his presentation at the 2004 Web-Wise meeting, suggests that, "There is as yet on the horizon no real substitute for the vision, discipline, and commitment needed to build digital collections at a scale and level of generality that will attract a broad audience of users and have such an impact on scholarship that their disappearance is not an option." Waters goes on explicitly to express his concern that the ad-hoc collection registries and metadata repositories over heterogeneous collections will not be adequate and sufficiently persistent to support scholarship over the long term. (Waters, 2004)

At the other extreme are proponents of services like Google, which assume the ad-hoc reusability of content, but (currently at least) are at best ambivalent as to the need for any special accommodation for reuse and repurposing, such as the creation of quality collection-level and item-level descriptive metadata. Google at present makes essentially no explicit use of manually-generated descriptive metadata or collection-level description, instead relying on brute computing force and free-text keyword indices and queries to provide search services over heterogeneous, disorganized full-content (or partial full-content). While there are indications of changes here – search engine system designers are showing renewed interested in metadata and are undertaking new initiatives to expose pieces of the “hidden web,” particularly those of interest to researchers, as recent collaborations between Google and DSpace and Yahoo and OAIster demonstrate (Young 2004, Suber 2004)) – this school of thought assumes that most traditional metadata paradigms are superseded by information retrieval operations over full-content and in any event do not scale well enough to be useful in the Web environment.

Clearly both of these contrary perspectives are correct in some contexts. The question is whether there exist contexts and real-world use cases that fall in a middle-ground niche between these two extremes. Are there information needs in practice that aren't met by Google-like approaches but for which large scale (and accordingly high-cost) monolithic digital library solutions of the sort envisioned by Waters would be overkill? Are there in fact information needs in practice that can be well enough met by services built over ad-hoc collection registries and item-level metadata aggregations? Through the implementation of generic formal and informal standards for sharing collection information and item-level descriptions, can communities of interest build effective and useful digital library services across distributed collections of digital content developed originally for diverse audiences and with diverse intended purposes? And is the time and effort spent on achieving this middle road capability worth it?

The answers are not immediately obvious. Our current effort is not of sufficient scope to fully answer these questions, but it is an explicit goal of the IMLS Digital Collections and Content (DCC) project [5] to make progress on this point and at least offer a contribution to furthering our understanding of the potential utility of general purpose collection-

level and item-level metadata in the implementation of search and discovery services across heterogeneous digital collections. By constructing and investigating the utility of a collection registry and metadata aggregation for IMLS NLG digital collections and content, we hope to provide at least anecdotal evidence pertinent to these issues. To accomplish these goals, as well as to meet the immediate needs of the IMLS NLG community for a shared content registry, we identified several intermediate project objectives prior to the start of our work a year and a half ago:

- Survey IMLS grantees to establish baseline of current practice, attitudes towards metadata, and technical readiness to implement OAI-PMH.
- Define collection-level metadata schema for collection registry; concurrently define initial models for searching & browsing of collection registry.
- Make available software and provide technical advice to encourage and facilitate grantee implementations of OAI-PMH.
- Implement working and updatable collection registry; target participation 90%.
- Harvest grantee metadata and implement search service across harvested aggregation of metadata; target participation 50%.
- Analyze quality and consistency of harvested item-level metadata from perspective of usefulness for interoperability.
- Investigate the research question, "How can resource developers best represent collections and items to meet the needs of service providers and end users?"
- Test usefulness of collection registry and item-level metadata aggregation with selected user populations.
- Report on observations and issues regarding barriers to interoperability, potential for useful and marketable digital library services built on ad-hoc collection registries and item-level repositories, and challenges and prerequisites for production implementations of registries and repositories.

Near the end of the project, based in part on our findings, the IMLS will decide whether to migrate prototypes of the collection registry and metadata repository we develop into permanent, production services.

The IMLS DCC project is a collaboration between the University of Illinois Library and the University of Illinois Graduate School of Library and Information Science (GSLIS). The focus of the Library project team is on the implementation of the collection registry and the item level metadata repository and draws on an extensive background in digital library infrastructure work, particularly in the OAI protocol. Timothy W. Cole, Mathematics Librarian, is the Principal Investigator (PI), and the co-PIs within the UIUC Library are William Mischo, Engineering Librarian, and Nuala Koetter, Interim Head of the Digital Media and Technology Initiative. The focus of the GSLIS project team is on the research question mentioned above: “How can resource developers best represent collections and items to meet the needs of service providers and end users?” To this end Associate Professors Carole Palmer and Michael Twidale, in conjunction with Research Assistants Ellen Knutson and Besiki Stvilia, are conducting interviews with IMLS NLG recipients and assessing metadata quality and use issues within the context of the local environment as well as within aggregations. This paper does not address the GSLIS research directly, but other preliminary reports on this work are available. (Knutson, Palmer, and Twidale, 2003; Palmer and Knutson, 2004)

Top-Level Architecture Decisions

To create working prototypes of an IMLS collection registry and item-level metadata repository, we needed to make early design decisions in two critical areas:

1. Selection of a model for cross-collection searching of item-level metadata.
2. Selection of a model for the collection registry, specifically what entities would be described and included in the collection registry.

Cross-collection searching of item-level metadata

The first decision, to use an OAI-PMH based metadata harvesting approach for collecting, aggregating, and searching item-level metadata, was explicitly required by the terms of the IMLS Request for Proposals (RFP) for this project, reflecting an assessment by IMLS that OAI-PMH is appropriate for project objectives and practicable and doable within project constraints. Based on our prior experience with OAI-PMH we concurred in this assessment. (Shreeves, Kaczmarek, & Cole, 2003)

From the perspective of IMLS, OAI-PMH offers a low-barrier approach to metadata sharing that is technically within reach of at least most NLG projects developing digital content. Both turnkey commercial and Open Source OAI-PMH solutions are available. Technical barriers have been further lowered with the recent addition of recommended

guidelines for OAI Static Repositories and Gateways, [6] and new work is now underway to create a module for Open Source Apache Web servers that would automatically export via OAI-PMH Web page metadata contained in HTML <meta> tags. [7] Though unfamiliar to some classes of cultural heritage institutions, the metadata harvesting model of OAI-PMH follows naturally from union catalog traditions in the library domain, and is conceptually congruent with the way large, well-established library cataloging utilities such as OCLC aggregate metadata for print content.

While broadcast search approaches by definition are not designed to aggregate metadata locally on a single server, such approaches can and have been used to create virtual metadata aggregations and support cross-collection searching (e.g., the initial Dienst / NCSTRL implementation which supported one-stop searching of metadata records describing university computer science reports issued by institutions from across the globe (Davis & Lagoze, 2000)). As compared to broadcast search approaches, the OAI-PMH harvesting and aggregation approach offers net pluses. Like OAI-PMH, broadcast search models assume widely distributed primary content, and most broadcast approaches rely on metadata for search and discovery. The primary difference is that broadcast search approaches rely on real-time, simultaneous processing of end-user search requests by all content providers sharing content. This approach is technically challenging for smaller content providers and does not scale well in heterogeneous computing environments as the number of participating content providers grows. Broadcast search is only as reliable as the least reliable content provider in the group. Often in broadcast search models there also is divergence in how search semantics are interpreted across a heterogeneous union of content providers. Metadata aggregation approaches like OAI-PMH allow the harvester to normalize and enrich metadata aggregated and helps insure a more uniform and consistent search across the full catalog of metadata being shared. Aggregators can more easily analyze the full body of metadata being made available, thereby providing useful and more complete feedback to content providers about the consistency and quality of their metadata (at least in terms of utility for interoperability). For all these reasons OAI-PMH made sense for this project as the preferred model for cross-collection searching of item-level metadata.

Collection registry model

Our design decisions for the collection registry centered on two distinct issues:

- What are the entities to be included in the collection registry?
- How will those entities be described?

We initially equated NLG projects with collections; that is, we assumed our NLG collection registry would simultaneously be a NLG project registry. This was an oversimplification. Closer investigation and input from the project steering committee made it clear that this approach led to confusion. How were we to deal with projects that involved multiple collections? How were we to deal with collections that were developed over the course of multiple projects? How were we to deal with situations where collection description attributes were not congruent with project description attributes – for instance, where the project administering institution was not the same as the collection owning institution? To resolve these issues we quickly moved towards a registry model which distinguished collections from projects. The primary entity in our registry is now explicitly the collection. Projects and other entities (e.g., related collections and agents) are maintained as separate entities and only described as necessary to establish their linkage and relation to a collection(s).

The decoupling in our registry scheme of NLG projects and the collections to which they are related represented an important (although perhaps obvious in retrospect) design decision. Often times a NLG project's primary goals are not the creation of a digital collection; instead they are training or collaboration or development of infrastructure. The digital collection created as a result of these activities is an important, but not fundamental end result of the project. Equating the digital collection with the IMLS NLG project would be misleading at best. An IMLS NLG Project Registry, though potentially a valuable resource and a recognized need within the IMLS community (such a registry was mentioned several times during the IMLS-sponsored "Digital Resources for Cultural Heritage: Current Status, Future Needs: A Strategic Assessment Workshop" held in August 2003) is not a goal for our project.

This brought us to the question of what is a collection within the context of an IMLS collection registry. That the collection was digital and was created or developed with at least some IMLS NLG funding are two givens. Beyond that the definition of 'collection' runs a wide gamut. Definitions vary from broad ('any aggregation of individual items', including an aggregation of one, based upon almost any criteria (Johnston & Robinson, 2002)) to specific (information environments which facilitate information seeking by providing a context for

resources selected and organized with a particular focus on the user (Lee 2000)). Our particular question is not unique. Hill et al. have documented the struggles of the Alexandria Digital Library team to define a digital collection. (Hill et al, 1999)

Following from this research and discussions within the project team, we determined some necessarily broad criteria for inclusion of collections in the registry. In addition to the requirements mentioned above, collections were also to be:

- Cohesive (whether by topic area, type of material, etc.);
- Searchable as a distinct collection;
- Available through a unique point of entry (i.e., a unique URL).

A collection could have multiple sub-collections, provided these meet the same criteria above.

The last criterion is largely practical and based upon the following user scenario. Imagine that a large collection has multiple sub-collections without distinct URLs. If a search retrieves several of these sub-collections but the entry is always to the same top-level URL, a user may not understand the distinction between these various sub-collections. Requiring a unique URL will aid in eliminating confusion of being directed to the same URL multiple times.

Once we had decided what to describe, we faced the natural follow-on: how to describe collections. We began by surveying what work had already been done on collection description. The use of systematic, standardized collection-level description, or collection-level metadata, for digital content is not very common in the United States except in the domain of archives, where the Encoded Archival Description (EAD) is used to mark up finding aids. Archival finding aids are what Heaney calls a 'hierarchic finding-aid,' that is, the collection description contains information about the collection as a whole as well as information about the individual items within the collection. Because the IMLS Digital Collections and Content project does not aim to describe both levels of description in a single registry, and because the creation of finding aids, whether in EAD or not, is a resource intensive enterprise, the use of EAD as the internal collection-level description schema of our registry was discarded as an option.

Much work, however, has been done in the United Kingdom on 'unitary finding aids' or collection-level description that contains only information about the collection as a whole and not about the individual items within it. The Research Support Libraries Programme (RSLP) Collection Description Schema (hereafter referred to as the RSLP CD schema) contains descriptive attributes about a collection, its location, agents associated with collection, and relationships with internal or external collections. [8] The RSLP CD schema is well documented and has been implemented – often with some modifications – by RSLP projects throughout the UK. [9] However, it has not been well tested for use in describing digital collections. The Dublin Core (DC) collection description application profile [10], currently in development, is based heavily on the RSLP CD schema, but has been adapted and somewhat simplified for digital collections. For instance, it does not attempt to describe the location(s) or agent(s) associated with a collection. We also spent some time examining the metadata schemas used in large, active collection registries such as Cornucopia, a database of museum collections in the UK [11], the National Science Digital Library [2], and EnrichUK [12], a registry of collections created through the New Opportunities Fund in the UK. After an analysis of these registries and schemas as well as discussions with the authors and maintainers of the RSLP CD schema, we determined that an adaptation of both the RLSP CD schema and the DC Collection Description Application Profile would best fit our needs. We discuss the further development of the IMLS DCC Collection Description Metadata Schema below.

Challenges Faced

Three significant challenges we encountered early on in the implementation of the IMLS collection registry and item-level metadata repository were:

1) The heterogeneity of the IMLS-funded digital collections and content;
2) Issues of metadata quality and consistency;
3) The wide range of readiness, willingness, and technical capabilities among the NLG projects for implementing the OAI protocol.

Heterogeneity of IMLS funded digital collections and content

When the Illinois team was awarded the grant, IMLS provided us with the grant proposals of all National Leadership Grants with digital content funded from 1998 through 2002. These proposals allowed us to document, at least to first-order, key characteristics over a range of 95 NLG projects. Specifically we used the proposals to identify institutions involved, project goals, collections created, content digitized or created, descriptive metadata schemes used, and technical specifications such as the content management system and whether an OAI data provider had already been planned or

implemented. This information was updated and supplemented through a survey distributed in September 2003 to 92 PI's representing 94 NLG projects with digital content [13]. We identified five non-active projects through the survey or other communications, leaving a survey population of 87. Our return rate was 76%. (This survey was sent to an additional 27 recipients of 2003 National Leadership Grants and non-respondents from the 1998-2002 pool in early May 2004; the information below refers only to the first round of the 1998-2002 NLG pool.)

The results of the survey and grant proposal analysis provide evidence of a diverse universe of IMLS funded digital collections and content. We found in particular:

- A wide diversity of institutions and collaborations
- Many different types of digital collections and sub-collections
- A broad range of item level metadata schemas and controlled vocabularies in use

Each of these characteristics of the population of collections and content for our project represents a challenge that must be addressed.

Diversity of institutions and collaborations

Of the 95 NLG grant proposals examined over half (54%) were collaborative efforts between multiple institutions. Including these collaborative partners we identified at least 237 distinct institutions from the grant proposals alone; however, after incorporating survey results and creating 84 preliminary collection registry entries, we actually documented 330 distinct institutions which have contributed to the digital collections described in our registry. Many of these contributors were not recorded on the grant proposal. The types of institutions range from large academic libraries with established digital library programs to small historical societies with little or no expertise in digital content creation. Figure 1 illustrates the types and numbers of institutions involved in the creation of the digital content as indicated from the grant proposals.

Take in Figure 1 here
Caption: Types of Institutions represented in NLG projects (from 1998-2002 grant proposals only)

The diversity of institutions – particularly within collaborative efforts – has an immediate, direct impact, as well as a more intangible impact, on our efforts. Our decision to enumerate each institution which contributes to a digital collection has meant that some collections, created through state-wide or broader collaborative efforts, are linked to literally a hundred or more institutions. Collections could potentially have many sub-collections organized by the contributing institution. These were considerations in the design of the database supporting the collection registry. At a more granular level the item-level metadata often points back to the institution hosting the aggregate digital collection, rather than the actual contributor. Although this is reliant on how the metadata is created and mapped at the data provider end, it impacts on how we might link institutions to the content they created or contributed to a larger collection.

The less tangible aspect of institutional diversity is in the world-view of the types of institutions represented here. Although all are broadly cultural heritage organizations, it is well recognized that museums, archives, and libraries each view the use and presentation of collections and content differently. In addition, although we began the paper speaking about cultural heritage, National Leadership Grants are also awarded to scientific organizations such as zoological societies and herbariums. In addition, NLG-funded projects often have specific uses in mind for the digital collections they create, such as use by the K-12 community or by specialists. These differences in perspectives directly impact how collections and particularly content are described.

Table 1 shows two metadata records exported in simple Dublin Core through the OAI protocol. Each describes separate instances of the same World War II poster. The first metadata record is from a large academic library and has been cataloged in a manner consistent with traditional library practice. The second metadata record is from a NLG project whose primary goal was to promote the use of digital content within the curriculum of elementary and middle school teachers through collaboration between the teachers and content creators. To this end the metadata includes interpretive information and learning standards (16 History, for example) to which the poster could belong. Aggregators of item-level metadata from diverse organizations have to find mechanisms to cope with metadata created for different use environments and identify metadata records describing duplicate or closely related information objects.

[Take in Table 1 here]
Caption: Comparison of metadata records describing separate instances of the same object

Types of Digital Collections and Sub-Collections

As we began to examine the output of each of the NLG projects and as we received the survey results, we found that although most NLG projects created

more or less traditional collections, albeit digital, a few of the collections were highly non-traditional, for instance: a multimedia exhibit that allows users to experience the oral history and visual images of a region simultaneously (Voices of the Colorado Plateau [14]), a web site that actively tracks wildlife conservation efforts in the field (Field Trip Earth [15]), and digital art projects ("Banana", "Code City", and "Hard Place" at the Lower East Side Tenement Museum [16]). Some of the more traditional digital collections included significant investment in peripheral material such as lesson plans, bibliographies, and contextual essays. The objects represented in these collections vary widely and include almost any type of material from manuscripts to maps to data sets to artifacts. Our challenge was to develop a collection-level metadata schema that could describe a wide range of these digital collections. Strategies we developed include the addition to the IMLS DCC Collection-level Description Schema of descriptive fields such as "Supplementary Materials."

Sub-collections also represent a potential challenge. Early in our discussions, we decided that we would allow the inclusion in our collection registry of records describing sub-collections at one level down from the parent collection (i.e. sub-collections could not have children). In order to gauge the number of sub-collections that might be created, we asked in the survey whether the respondents had sub-collections, how many, and how these were organized. 76% of the respondents reported that their collection was divided into sub-collections. 38% reported that they had between 2-5 sub-collections while 22% reported that they had 6-10 sub-collections. Interestingly, a handful of respondents reported having many hundreds of sub-collections. In these cases the division was based on the subject headings used; every subject heading represented a distinct sub-collection. Table 2 reports the organization of sub-collections. Note that 36% of the respondents reported organizing sub-collections on the basis of two or more factors.

[Take in Table 2 here]
Caption: Basis of sub-collection organization (results from survey of 1998-2002 NLG recipients)

The challenge here is two-fold. Again the collection-level description metadata schema used must be robust enough to handle a variety of descriptions. The RSLP CD schema and the Dublin Core Collection Description Application Profile have proven generally satisfactory in this regard, though we did have to make a few small customizations. [17] The structure for the database must also handle a proliferation of sub-collections, and the registry display must communicate these structures and relationships to the user. This last requirement is especially difficult and we are still working on ways to satisfy this need.

Item-level metadata schemas and controlled vocabularies in use

Eighty-six percent (86%) of the survey respondents reported using item-level metadata to describe the resources within their collections. The metadata standards most often in use are Dublin Core (56% of respondents with item level metadata) and MARC (33% of respondents with item level metadata). Other standards used include EAD, the Text Encoding Initiative (TEI) Header, Visual Resources Association (VRA) Core, Darwin Core, Making of America (MOA) 2, and the Taxonomic Data Working Group-Structure for Descriptive Data (TDWG-SDD). The diversity of item-level metadata in use by NLG projects is not surprising. Perhaps what is surprising are the number of respondents with item-level metadata who use locally developed schemas (39%) and the number who use multiple schemas (61%). Figure 2 illustrates the diversity of metadata standards (and non-standards in use).

[Take in Figure 2 here]
Caption**:** Metadata schemas in use (results from survey of 1998-2002 NLG recipients)

It should be noted that not all of the digital content created through the National Leadership Grant program has item level metadata. 14% of the respondents reported not using descriptive metadata to describe the contents in their digital collection. These respondents for the most part had created collections that are not easily divisible into discrete items, such as multimedia exhibits, learning objects, or heavily integrated web pages, and who provide no search services for specific individual resources. We cannot, of course, include these collections in the item-level metadata repository, although they will be represented in the collection registry.

The diversity of metadata schemas can pose a significant challenge for the implementation of OAI data provider services. The OAI protocol requires the provision of metadata in at least simple Dublin Core. In order to implement OAI data provider services, NLG projects need to map their native metadata schemas to simple Dublin Core. Cross-walking between metadata schemas is not a trivial process and can be a barrier to implementation as many organizations are understandably reluctant to lose the complexity and semantic structure of their chosen metadata schema to the bluntness of Dublin Core.

OAI-PMH supports the use of metadata schemas in addition to Dublin Core, and we continue to encourage implementers of OAI data provider services to provide their metadata in its native schema as well. This does, however, require (for validation purposes) that content providers implement or point to valid and correct XML schemas for all metadata formats other than Dublin Core that they export. Locating or creating such XML schemas is not necessarily a simple task, particularly when working from a unique, local metadata schema.

Eighty-four percent (84%) of the respondents with item level metadata reported using some form of controlled vocabulary in their item level metadata. Table 3 identifies the most used controlled vocabularies for five types of values: subject, format, type, personal names, and geographic names.

[Take in Table 3 here]
Caption: Most used controlled vocabularies for five value types (results from survey of 1998-2002 NLG recipients)

The diversity of controlled vocabularies and metadata schemas complicates the creation of an effective item-level metadata aggregation and has an impact on the utility of metadata for interoperability. This impact is additive to the impact of metadata quality and consistency generally within the sets of records contributed by each participating repository, as discussed in the next section.

Metadata Quality in Aggregate

The challenges that face OAI service providers (metadata harvesters) when aggregating metadata from multiple data providers are well documented (Shreeves, Kaczmarek, & Cole 2003, Halbert 2003, Arms et al 2003), and those facing the IMLS item-level metadata repository are no different. Briefly, some of the aggregation issues include:

- Disparate and inconsistent use of Dublin Core elements.
- Loss of information when providers map from more complex and expressive metadata schemas to simple Dublin Core.
- Loss of browse capabilities due to diversity of controlled vocabularies and encoding schemes being used.
- Varying practice in granularity of description and distinctions about what is described (e.g., the physical artifact photographed or the digital manifestation/surrogate of the physical artifact).
- Variations due to broad range of types of material described.

What these issues illustrate is that the OAI community has yet to come to grips with what quality "shareable" metadata is. While there has certainly been work done on best practices for specific communities or domains (The Western States Dublin Core Metadata Best Practices [18] and the Open Language Archive Community [19] are two examples), there has been little research into what are the key attributes or metrics of quality for "shareable" or "interoperable" metadata. It may be that metadata of high quality within a local context is of significantly lesser quality (at least in terms of utility) when taken out of its local context and aggregated with other metadata records. Just as libraries had to come to grips with these sorts of interoperable quality issues when MARC records were shared via OCLC (Maciuszko 1984) so too does the digital library community need to address these issues in the age of federated digital content.

The GSLIS research team is currently investigating how to measure and assure metadata quality in aggregated digital collections. They are empirically examining the harvested metadata to develop systematic techniques for metadata quality assessment and assurance. We anticipate that this work not only will help content providers create metadata more useful in a shared context, but also will suggest ways in which OAI service providers can better normalize and/or enrich aggregated collections of metadata.

OAI Capability and Readiness

Based on proposal analysis and the results of our initial survey, Table 4 gives a preliminary assessment of the capability of the original pool of 94 1998-2002 NLG projects to implement OAI data provider services. While 44% of NLG projects either have or actively plan to implement the OAI protocol, 20% of respondents to the survey indicated that they had not heard of the OAI protocol.

[Take in Table 4 here]
Caption: Breakdown of NLG recipients according to readiness to implement OAI data provider services

Beyond marketing the capabilities and potential of OAI through tutorials, presentations, and one-on-one conversations, we are also tracking why NLG projects might not be able or ready to implement data provider services. A preliminary review indicates that NLG projects may not be in a position to implement OAI data provider services for any of several reasons:

- There is no item level metadata. This is true for many exhibit and learning object focused projects.
- The collection is not yet public. NLG projects wish to wait until they unveil their digital collection before sharing the metadata.
- Infrastructure is not in place. The metadata may not be mapped into Dublin Core or stored in a manner to easily support implementation of OAI data provider services. Necessary technical expertise may not be available. (Obviously these are especially problems for projects that have fully expended their NLG grants and may have no other available resources to implement infrastructure enhancements.)
- The technical infrastructure is in transition or will be in transition. NLG projects are reluctant to implement OAI provider services in the midst of a migration to a new content management system.
- Agreement has not yet been reached among all collaborators of a specific project to share that project's metadata via OAI.

Some of these barriers are insurmountable (we can't harvest item-level metadata if there is none!), but we are actively working on ways to facilitate implementation of OAI by NLG grantees in other instances.

Accomplishments to Date

At this, the mid-point of the IMLS Digital Collections and Content project, we have made progress on several fronts. Accomplishments to date are listed here and discussed in more detail below:

- Creation of the IMLS DCC Collection Description Metadata Schema;
- Development of a beta IMLS Digital Collection Registry and Registry Entry/Edit Forms;
- Facilitation of implementation of OAI data providers for NLG funded collections;
- Development of a beta repository for item-level metadata harvested from NLG funded digital collections.

IMLS DCC Collection Description Metadata Schema

The IMLS DCC Collection Description Metadata Schema [17], as mentioned above, is based on the RSLP Collection Description Metadata Schema and the Dublin Core Collection Description Application Profile. The IMLS DCC project has adapted these schemas to reflect the particular nature of the project and to incorporate the specific needs of our NLG collection registry. The resulting schema is meant to describe the digital collections created through IMLS funded NLG projects and does not describe in any detail the projects themselves. This metadata schema forms the basis of the IMLS NLG Collection Registry, currently in beta phase of development.

The following is meant only to give a cursory overview of the schema. There are four classes of entities described by the schema:

- collections, including both NLG collections and physical or digital collections associated with (related to) a NLG collection;
- NLG projects associated with a NLG collection;
- institution(s) associated with a collection and/or a NLG project; and
- administrators of NLG collections.

A collection may have been created by multiple NLG projects and have multiple administrators. A collection may have only one hosting institution, but may have multiple contributing institutions. A collection may have multiple sub-collections, associated collections, or source physical collections. A NLG project may have only one administering institution, but may have multiple participating (or collaborating) institutions. Figure 3 below illustrates the relationships between these entities. The complete list of schema elements (i.e., entity descriptive attributes) is available on the project Website. An XML schema definition (.xsd) file appropriate for validating collection description metadata records is currently being finalized and will be added to the Website soon.

[Take in Figure 3 here]
Caption: Relationships between Entities in the IMLS DCC Collection Description Metadata Schema

IMLS Digital Collection Registry

As mentioned above several existing, active collection registries were examined for functionality, interface design, and metadata schema. In January 2003, we consulted with David Dawson of Resource: The Council for Museums, Archives & Libraries (now MLA) in the UK about Cornucopia and plans to develop a registry for the NOF-Digitise project (UK) and the Minerva project (Europe). Through examination of these registries and conversations with David and others, we identified several functions to be included in the IMLS DCC registry. They include browsing by topic area, expressing

relationships among collections (e.g. parent-child), and limiting searches by time period, geographic area, audience, and/or type of material. We also wanted the NLG projects to be able to edit their own collection registry records, so we examined several collection registry input forms such as those used by the NSDL and RSLP for design and functionality issues.

Much of our development work thus far has been to design and test the database for the collection registry records and design the registry entry/edit forms. We are currently in the last stages of iterative design of the registry entry/edit forms. In the winter of 2003/04 84 collection records were created from the survey results, then edited and expanded through information gleaned from collection websites and other communications. A preliminary browse interface for the collection registry has been developed as well. This, however, will undergo several more iterations. A staff view (partial) of a collection registry record is given in Figure 4.

[Take in figure 4 here.]
Caption: Browser screen showing partial record from IMLS DCC collection registry (beta version)

Facilitating Implementation of OAI Data Providers

We have pursued several strategies for facilitating OAI-PMH implementation by IMLS NLG projects interested in participating in our IMLS DCC item-level aggregated metadata repository. While project funded has generally precluded on-site visits to implement OAI-PMH software, we have been able in a few cases to customize existing Open Source software solutions for use in the specific grantee environments. In other cases we have been able to assist by exercising (testing) and vetting OAI-PMH implementations created by grantees. Often implementations in these latter cases have been commercial turnkey solutions. Our testing has helped identify bugs or other possible issues or concerns with implementation of those solutions in specific grantee environments. A few examples are given here to illustrate the nature of this phase of our activities.

In February 2003, we completed remote installation of an OAI data provider service for the Colorado Digitization Program (CDP). This service was implemented on top of the metadata storage infrastructure already in use by CDP, but did not require any changes by them to existing metadata processing or workflow. The CDP service supports exporting metadata in a qualified Dublin Core schema, as well as in simple Dublin Core. The implementation took advantage of pre-existing Apache Web server and MySQL implementations in place on the CDP servers. Tomcat extensions were added to the Apache application to allow the implementation of the Java Serverlets that implement the actual OAI metadata provider protocol services. We customized an existing generic Java Serverlet Open Source OAI provider application we had previously developed on an earlier project. (Generic versions of the all University of Illinois Library metadata provider implementations and associated XML schema definitions created as part of this work are available on SourceForge under UIUC/NCSA Open Source licensing. [20])

In July 2003 we set up an OAI Static Repository [6] for the NLG project “American Natural Science in the First Half of the Nineteenth Century” based at the Academy of Natural Science. A recent development in the OAI protocol and designed for use with small, relatively static metadata collections, a static OAI repository is a single XML file which contains all repository metadata records and which sits on the metadata provider’s existing Web server. A third party acts as a gateway through which an OAI service provider can then harvest individual metadata records contained in that static XML file. This obviates the need for the source metadata provider to implement a dynamic web service. The project team worked with Eileen Mathias at the Academy of Natural Science to map metadata from MARC records to simple DC and produce a single XML file (with both MARC and DC records available for harvest) which is now available through an OAI Static Repository Gateway running on our servers at Illinois. The success of this implementation indicates that the static provider service approach is a good solution for institutions lacking technical infrastructure to implement new, dynamic web services.

In July 2003 we worked with the Washington State Libraries to test harvest metadata from their CONTENTdm data provider service. CONTENTdm is a digital library management system which has built in an OAI data provider service. At that time the then current version of CONTENTdm (3.5) did not support resumption tokens. These are an optional feature in the 2.0 OAI protocol which aid in 'flow control' by allowing a data provider to issue records in manageable chunks to a service provider, thus limiting the peak load on both systems. Although optional, the implementation of resumption tokens is particularly important for large data providers. The Washington State Library repository proved too large to function reliably without resumption tokens. We examined other possible avenues for harvesting these records. We determined that dividing metadata into smaller sets (maximum of 10,000 records per set) could facilitate harvesting without flow control. We

also developed a successful workaround in which we harvested records individually (using the OAI-PMH GetRecord verb instead of the more typically used ListRecords verb). While this work-around was slow, it put little to no stress on the web server and all metadata records were harvested successfully. However, based in part from feedback from us and Washington State, CONTENTdm has since implemented resumption tokens in their OAI provider module, improving robustness for large repositories using that software

Lastly, we created an OAI data provider for the IMLS-funded Illinois Alive project. The Illinois Alive collection consists of a series of web pages about Illinois history. Dublin Core metadata for each web page is embedded in the HTML Head element of each page. The IMLS DCC team developed a spider that crawled through the Illinois Alive pages to collect the Dublin Core metadata and store it within a SQL database on one of ourervers. The metadata is then exposed via the OAI protocol. Similar functionality (implemented in a simpler and more robust fashion) is expected from the in-progress project mentioned above to create an Open Source OAI-PMH extension module for Apache Web servers.

Item-Level Metadata Repository

To date 87,537 item-level metadata records have been harvested from 20 IMLS NLG collections using the OAI-PMH. Initial harvesting has been done exclusively in Dublin Core metadata format. Harvested records have been indexed in a Microsoft SQL database and a preliminary, early beta version of a Web interface has been implemented to allow searching of the metadata aggregation. An illustrative search result screen from this preliminary interface is shown in Figure 5.

[Take in Figure 5]
Caption: Sample result list from simple search of IMLS DCC metadata repository (beta version).

Repositories are revisited every three weeks for incremental harvesting, and once every three months full re-harvests are done of each repository. Periodic full re-harvests are required since most OAI repositories from which we are harvesting for this project do not support the optional feature of the OAI protocol requiring providers to maintain in perpetuity a record of all metadata items ever deleted from their repository. To this point in the project little normalization or augmentation of metadata records harvested has been done. Based on our preliminary inspection of metadata so far harvested we have identified several automated normalization and augmentation functions that will be implemented soon. Some normalization and augmentation will need to be done on a repository-by-repository basis, and some can be applied across the entire aggregation. We anticipate that the systematic analysis of metadata quality and consistency currently being performed by our GSLIS colleagues will suggest additional normalization and augmentation functions.

We also anticipate that the output of metadata normalization and augmentation processes will need to be stored (and indexed), internally at least, in a more expressive metadata schema than simple Dublin Core. We are currently testing the use of a qualified Dublin Core schema, extended with the addition of project-specific encoding and refinement semantics. This approach will allow us to harvest and take fuller advantage of optional richer metadata formats made available by some of the participating OAI metadata providers. Cross-walks from these formats to qualified Dublin Core will be less lossy.

Conclusions

The advent of the Web and other related digital technologies presents a good opportunity for increased content sharing and collaboration in the development of information systems. While a measure of interoperability, e.g., sharing generic HTML Web resources via Google, has proven relatively easy to accomplish, search and discovery across aggregations of more varied and complex digital content in a robust and full featured manner is proving harder than initially perceived by many of us. Making specialized scholarly digital content – primary content that is frequently non-textual, often hidden within complex database structures and collection contexts – more visible and easily accessible requires higher precision search and discovery systems that can exploit richer and more highly structured metadata. Issues of granularity and context are proving especially important when dealing with aggregation of such content.

It is not yet clear whether ad hoc collection registries and item-level metadata aggregations built using a generic metadata harvesting protocol such as OAI-PMH are sufficient to implement the next generation of cross-repository digital library search and discovery services. As described above, a number of challenges exist, even in the context of our relatively controlled experiment with IMLS NLG digital collections and content. Based on our experience so far, part of the problem appears to be a lack of clear guidance and well-established best practices, not for creating metadata generally, but for creating metadata optimized for aggregation and interoperability. Our project and several similar

projects currently in progress will help the community address this need. New metrics for metadata quality as defined in this context are emerging (Bruce and Hillmann, 2004), and at the very least we hope to help establish benchmarks for current metadata authoring practice and the implications of state-of-the-art practices for metadata harvesting and aggregation services.

A further goal, and one that we have borne in mind as we develop both the collection registry and the item-level metadata repository, is to link the two so that users can move between one and the other. The lack of context for any given individual resource in an aggregation could perhaps be mitigated by the effective delivery and integration of collection level description for that resource with its item-level description.

Finally, in the next phase of work on our IMLS DCC project, we hope to develop preliminary anecdotal evidence as to the potential benefit and utility of these kinds of services for one or two specific user populations. While a full-blown user study and analysis is beyond the scope of our current grant, we do plan during the final year of the project further small-scale user focus groups, usability experimentation, and transaction log analysis, building on early work in this vein on an earlier OAI-PMH based metadata harvesting service project. (Shreeves and Kirkham, in press)

Notes

[1] Institute of Museum and Library Services: http://www.imls.gov/.
[2] National Science Digital Library: http://www.nsdl.org/.
[3] National Information Standards Organization: http://www.niso.org/.
[4] Open Archives Initiative: http://www.openarchives.org/.
[5] IMLS Digital Collections and Content: http://imlsdcc.grainger.uiuc.edu/.
[6] Specification for an OAI Static Repository and an OAI Static Repository Gateway: http://www.openarchives.org/OAI/2.0/guidelines-static-repository.htm.
[7] Mod_OAI Project Homepage: http://www.modoai.org/
[8] RSLP Collection Description Schema: http://www.ukoln.ac.uk/metadata/rslp/schema/.
[9] For examples of projects that have implemented the RSLP Collection Description Schema see: http://www.ukoln.ac.uk/cd-focus/cdfocus-tutorial/lookslike.html.
[10] Dublin Core Collection Description Application Profile Summary: http://www.ukoln.ac.uk/metadata/dcmi/collection-ap-summary/.
[11] Cornucopia: Discovering UK Collections: http://www.cornucopia.org.uk/.
[12] EnrichUK: http://www.enrichuk.net/.
[13] Of the original 95 projects, two of the NLG projects were follow-ons to previous NLG grants. We only sent one survey in these cases. One NLG award was returned due to the dissolution of the institution receiving it.
[14]Voices of the Colorado Plateau: http://archive.li.suu.edu/voices/.
[15] Field Trip Earth: http:www.fieldtripearth.org/.
[16] See “Banana”, “Code City”, and “Hard Place” on the Lower East Side Tenement Museum website: http://www.tenement.org/features.html.
[17] IMLS DCC Collection Description Metadata Schema: http://imlsdcc.grainger.uiuc.edu/CDschema_overview.htm.
[18] Western States Dublin Core Metadata Best Practices: http://www.cdpheritage.org/resource/metadata/wsdcmbp/index.html.
[19] Open Language Archives Community: http://www.language-archives.org/.
[20] UIUC Open Source OAI Metadata Harvesting Project on SourceForge: http://uilib-oai.sourceforge.net/

References

Arms, W.Y., N. Dushay, D. Fulker, and C. Lagoze. (2003) “A case study in metadata harvesting: the NSDL.” *Library Hi Tech* Vol 21, No. 2, pp.228-237.

Bruce, T.R. and D.I. Hillmann. (In press 2004) “The continuum of metadata quality: defining, expressing, exploiting,” in *Metadata in Practice*, Ed. by Diane Hillmann and Elaine Westbrooks. Chicago: ALA Editions.

Cole, T. W. (2002) “Creating a Framework of Guidance for Building Good Digital Collections,” *First Monday* Vol. 7 No. 5. Available (May 13, 2004) at: http://www.firstmonday.org/issues/issue7_5/cole/index.html

Davis, J.R. and C. Lagoze. (2000) “NCSTRL: Design and deployment of a globally distributed digital library,” *Journal of the American Society for Information Science* Vol. 51 No. 3, pp. 273-280.

Dempsey, L. (2003) "The recombinant library: portals and people," *Journal of Library Administration.* Vol. 39 No. 4, pp.103-136.

Halbert, M. (2003) "The Metascholar Initiative: AmericanSouth.Org and MetaArchive.Org." *Library Hi Tech* Vol. 21 No. 2, pp.182-198.

Heaney, M. (2000) An Analytical Model of Collections and their Catalogues. Available (May 17, 2004) at: http://www.ukoln.ac.uk/metadata/rslp/model/.

Hill, L.L. et al. (1999) "Collection metadata solutions for digital library applications," *Journal of the American Society for Information Science* Vol. 50 No. 13, pp. 1169-1181.

Institute of Museum and Library Services. (2001) Report of the IMLS Digital Library Forum on the National Science Digital Library Program. Available (May 13, 2004) at: http://www.imls.gov/pubs/natscidiglibrary.htm

Johnston, P. and B. Robinson. (2002) "Collections and Collection Description." *Collection Description Focus Briefing Paper* No. 1. Available (May 17, 2004) at: http://www.ukoln.ac.uk/cd_focus/briefings/bp1/bp1.pdf.

Knutson, E., C. Palmer, and M. Twidale. (2003) "Tracking metadata use for digital collections [Poster Abstract]," in *DC-2003: Proceedings of the International DC Metadata Conference and Workshop*, pp. 243-244. Available (May 17, 2004) at: http://www.siderean.com/dc2003/706_Poster49-color.pdf.

Lee, H. (2000) "What is a collection?" *Journal of the American Society for Information Science* Vol. 51 No. 12, pp. 1106-1113.

Lynch, Clifford. (2002) "Digital Collections, Digital Libraries, and Digitization of Cultural Heritage Information," First Monday Vol. 7 No. 5. Available (June 1, 2004) at: http://www.firstmonday.org/issues/issue7_5/lynch/index.html on

Lyman, P. and H.R. Varian. (2003) How much information. Available (May 13, 2004) at: http://www.sims.berkeley.edu/how-much-info-2003.

Maciuszko, K. M. (1984) *OCLC: A Decade of Development 1967-1977*. Libraries Unlimited, Inc: Littleton, CO.

National Information Standards Organization. (2004) A Framework of Guidance for Building Good Digital Collections. Available (May 14, 2004) at: http://www.niso.org/framework/forumframework.html on.

Palmer, C. and E. Knutson. (2004) "Metadata practices and implications for federated collections," accepted for 2004 American Society for Information Science and Technology Conference.

Seaman, D. (2003) "Deep sharing: a case for the federated digital library," *EDUCAUSE Review* Vol. 38 No. 4, pp.10-11.

Sherman, C. and G. Price. (2003) "The invisible web: uncovering sources search engines can't see." *Library Trends* Vol. 52 No. 2, pp.282-298.

Shreeves, S.L. and T.W. Cole. (2003) "Developing a collection registry for IMLS digital collections [Poster Abstract]," in *DC-2003: Proceedings of the International DCMI Metadata Conference and Workshop*, pp. 241-242. Available (May 17, 2004) at: http://www.siderean.com/dc2003/705_Poster43.pdf .

Shreeves, S.L., J.S. Kaczmarek, and T.W. Cole. (2003) "Harvesting cultural heritage metadata using the OAI protocol," *Library Hi Tech* Vol. 21 No. 2, pp. 159-169.

Shreeves, S.L. & C.M. Kirkham. (In press) "Experiences of Educators Using a Portal of aggregated Metadata," *Journal of Digital Information* Vol. 5 No. 2.

Suber, P. (2004) "The case for OAI in the age of Google." *SPARC Open Access Newsletter* 73. Available (May 21, 2004) at: http://www.earlham.edu/~peters/fos/newsletter/05-03-04.htm.

Waters, D.J. (2004) "Building on success, forging new ground: The question of sustainability," First Monday Vol. 9 No. 5. Available (May 17, 2004) at: http://www.firstmonday.org/issues/issue9_5/waters/index.html.

Young, J. R. (2004) “Libraries aim to widen Google’s eyes: Search engines want to make scholarly work more visible on the Web.” *The Chronicle of Higher Education* Vol. 50 No. 37, p.A1.

Post-Print (final draft post-refereeing)
Published article: Cole, T.W. & Shreeves, S.L. (2004). Search and discovery across collections: The IMLS Digital Collections and Content Project. Library Hi Tech 22(3): 307-322 DOI: 10.1108/07378830410560107.

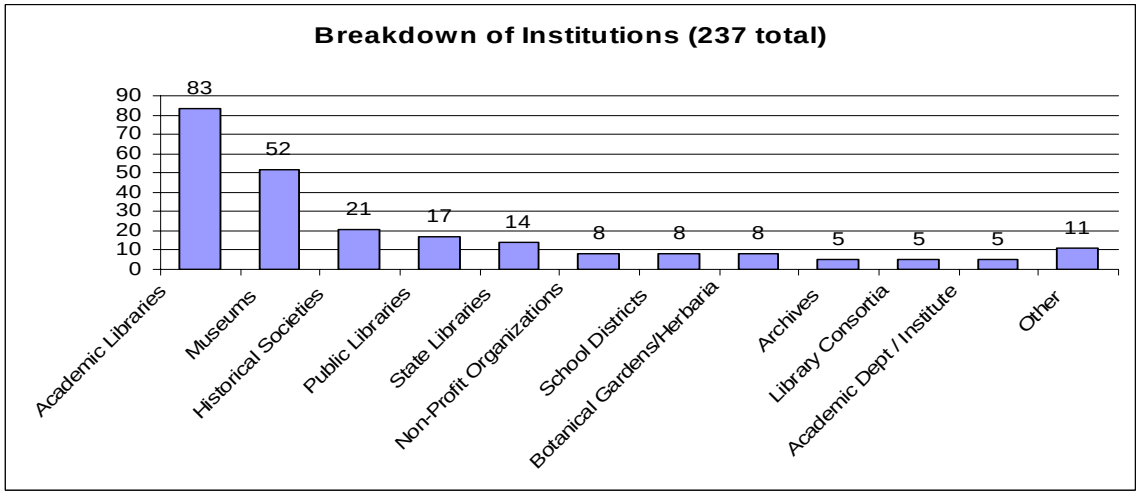

Figure 1 - Types of Institutions represented in NLG projects (from 1998-2002 grant proposals only)

Dublin Core Elements	**Record One (Traditional Library Cataloging)**	**Record Two (Access for Educators / Students)**
Title	Wanted! For murder: her careless talk costs lives	Wanted! For murder: her careless talk costs lives
Author	Keppler, Victor	*Not used*
Subject	— World War, 1939-1945. United States. — Espionage	World War II; War posters, American; National security; World War, 1939-1945 – Social Aspects – United States
Description	— "U.S. Government Printing Office: 1944—O-595600" — Woman's photograph. Poster promotes vigilance.	— Poster, b/with 27.9 x 20 in, published by the United States Government Printing Office. — During wartime concerns with about national security increase and World War II was no exception. This poster reminds citizens that sharing any military information such as troop movements, or other details could help the enemy sabotage the war effort. — World War II — 16 History; 14 Political Systems
Coverage	*Not used*	1944
Date	1944	3-22-02
Rights	Subject to U.S. and international copyright laws. Please contact the owning repository.	http://images.library.uiuc.edu/projects/tdc/conditions.htm
Language	Eng	*Not used*
Contributor	United States. Office of War Information.	*Not used*
Type	Poster	Image
Format	Image/jpeg	*Not used*

Table 1 – Comparison of metadata records describing separate instances of the same object

Basis of sub-collection organization:	**Number (%) of respondents with sub-collections:**
Administrative unit only	6 (12%)
Topic only	10 (20%)
Type of material only	8 (16%)
Other basis only*	8 (16%)
Based on two factors:	
Administrative unit and Topic	2 (4%)
Administrative unit and Type of material	1 (2%)
Administrative unit and Other	4 (8%)
Topic and Type of material	5 (10%)
Topic and Other	2 (4%)
Based on three factors:	
Topic, Type of material, and Administrative unit	4 (8%)
*Other responses included: Learning Standards; Grade level appropriateness; Keywords; Time-period; Audience; Donating individual or organization	

Table 2 - Basis of sub-collection organization (results from survey of 1998-2002 NLG recipients)

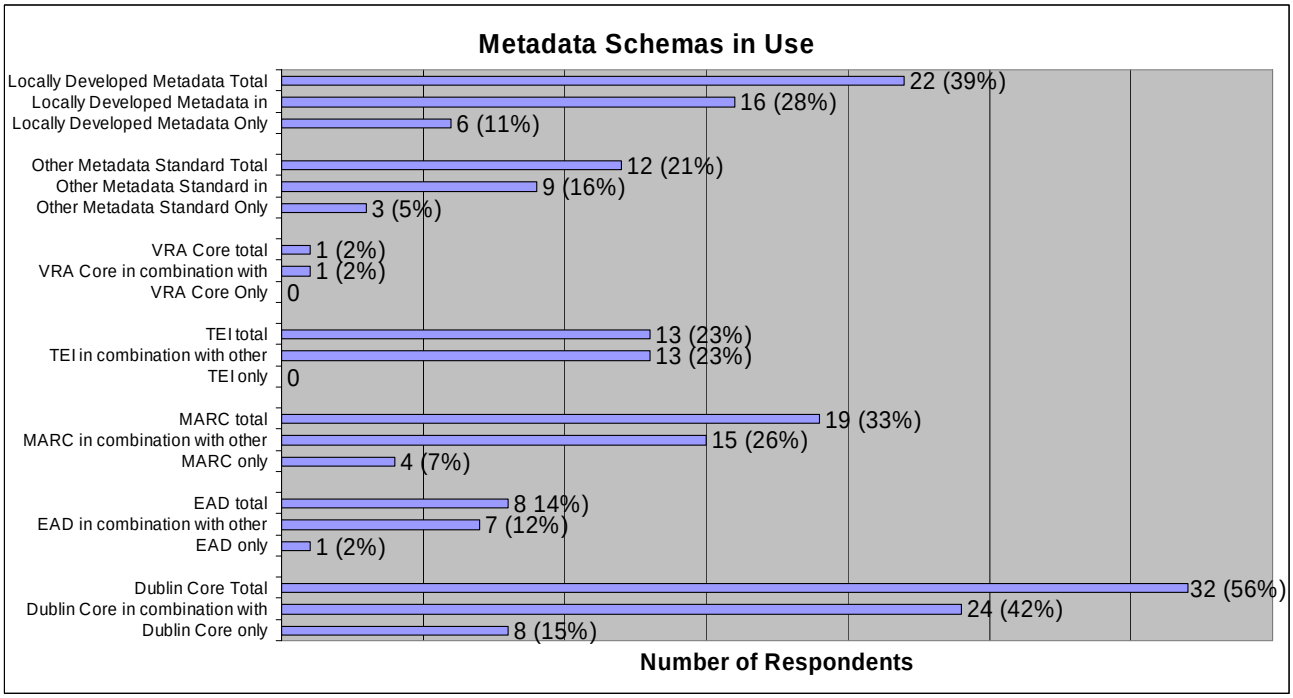

Figure 2 - Metadata schemas in use (results from survey of 1998-2002 NLG recipients)

Element	**Top three used Controlled Vocabulary (% of respondents who identified C.V.)**
Subject	LCSH (73%); LC TGM I (27%); AAT (17%)
Format	LC TGM II (17%); AAT (10%); MIME types (8%); AACR2 (8%)
Type	LC TGM II (21%); DCMI Type (13%); AACR2 (10%)
Personal names	LC Name Authority File (67%)
Geographic names	LCSH (27%); LC Name Authority File (25%); Getty Thesaurus of Geographic Names (15%)

Table 3 - Most used controlled vocabularies for five value types (results from survey of 1998-2002 NLG recipients)

Category of 1998-2002 NLG Recipients:	**Number / % of NLG Projects:**
Group 1 – Projects with OAI data provider sites for NLG content	21 (22 %)
Group 2 – Projects whose institutions have an OAI implementation (not yet being used for NLG content) or projects that have explicitly expressed plans to add OAI functionality	21 (22 %)
Group 3 – Projects who meet certain technical criteria – e.g. have item-level metadata and a maintained web site	23 (24 %)
Group 4 – Projects with no item-level metadata, no interest in providing metadata via OAI, or whose grants were given up	13 (14 %)
Unknown	17 (18 %)
Total	95

Table 4 - Breakdown of NLG recipients according to readiness to implement OAI data provider services

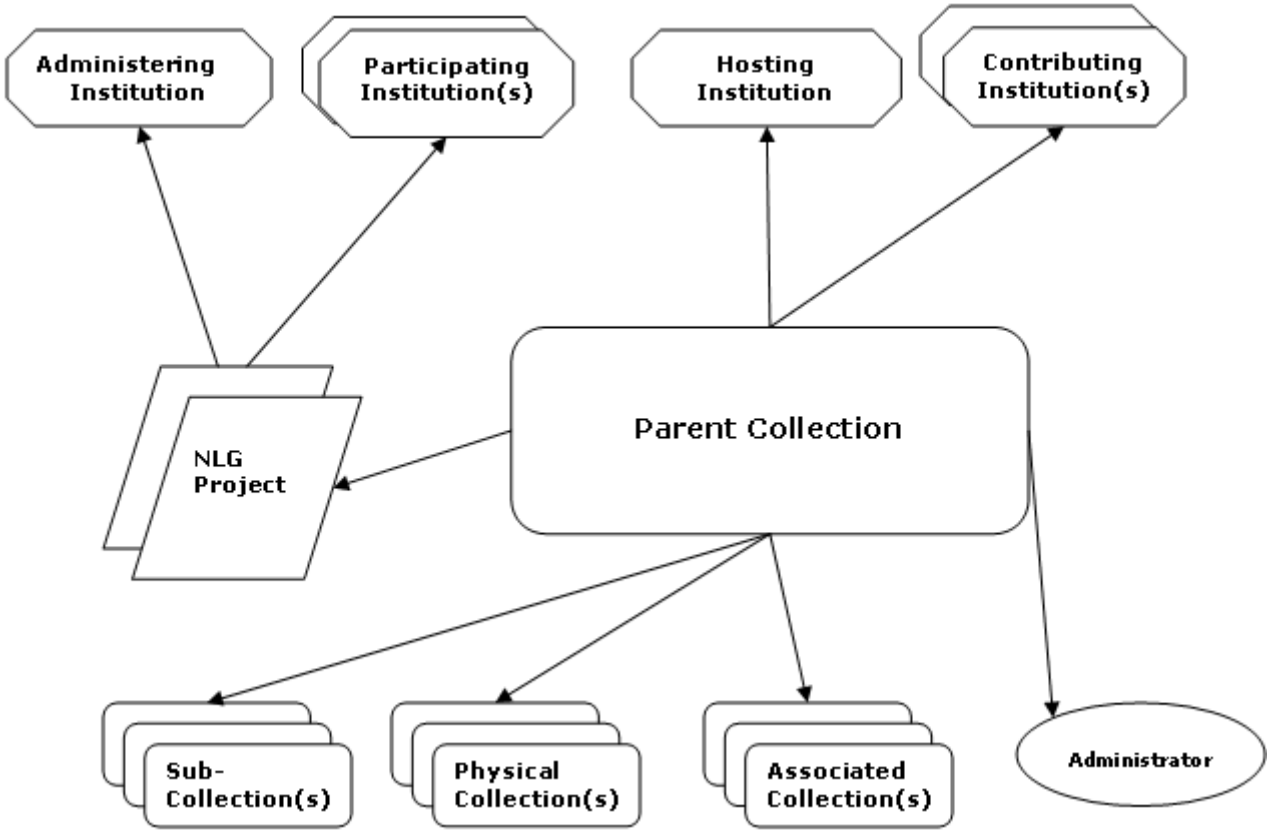

Figure 3 - Relationships between Entities in the IMLS DCC Collection Description Metadata Schema

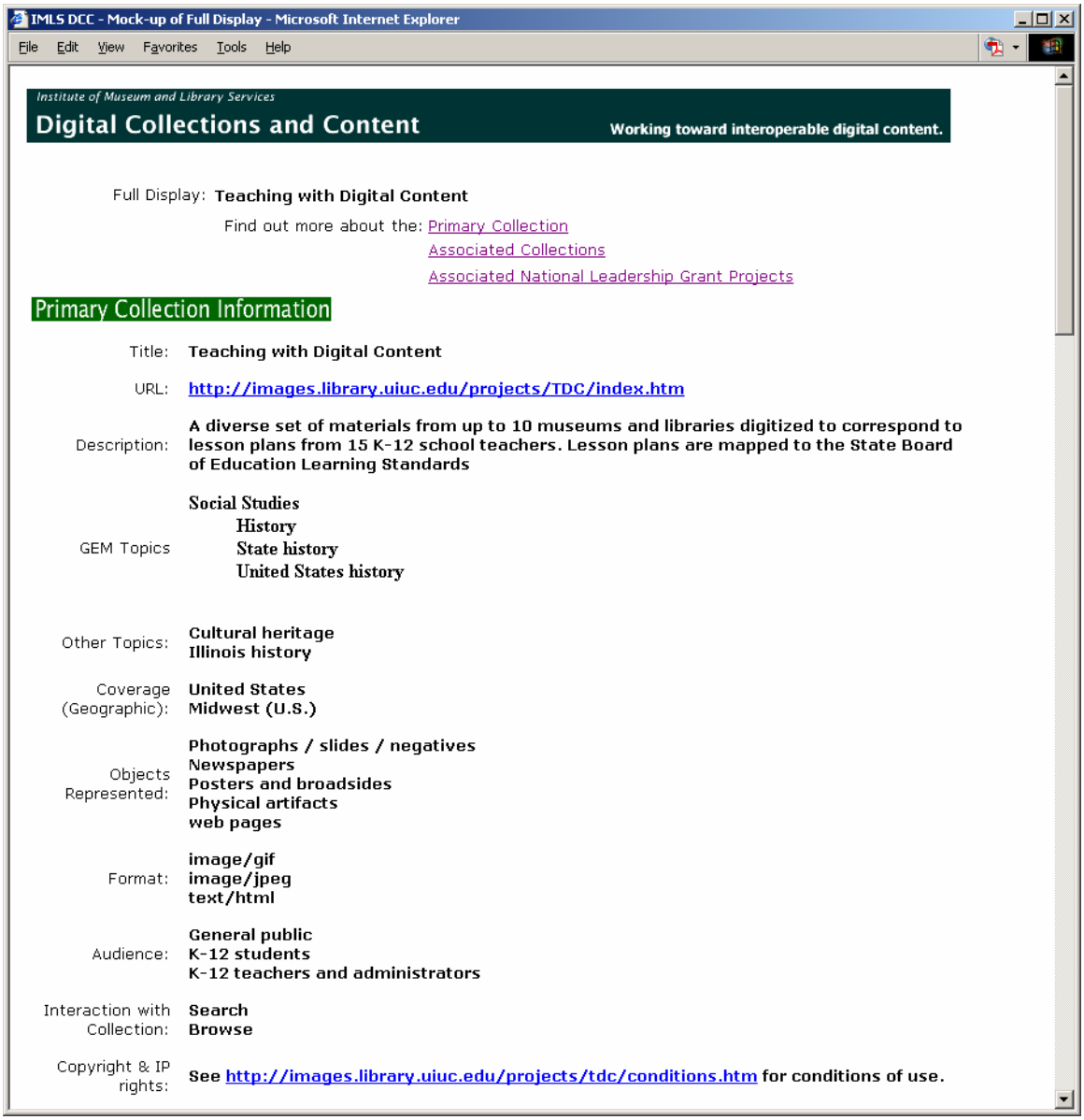

Figure 4 – Browser screen showing partial record from IMLS DCC collection registry (beta version).

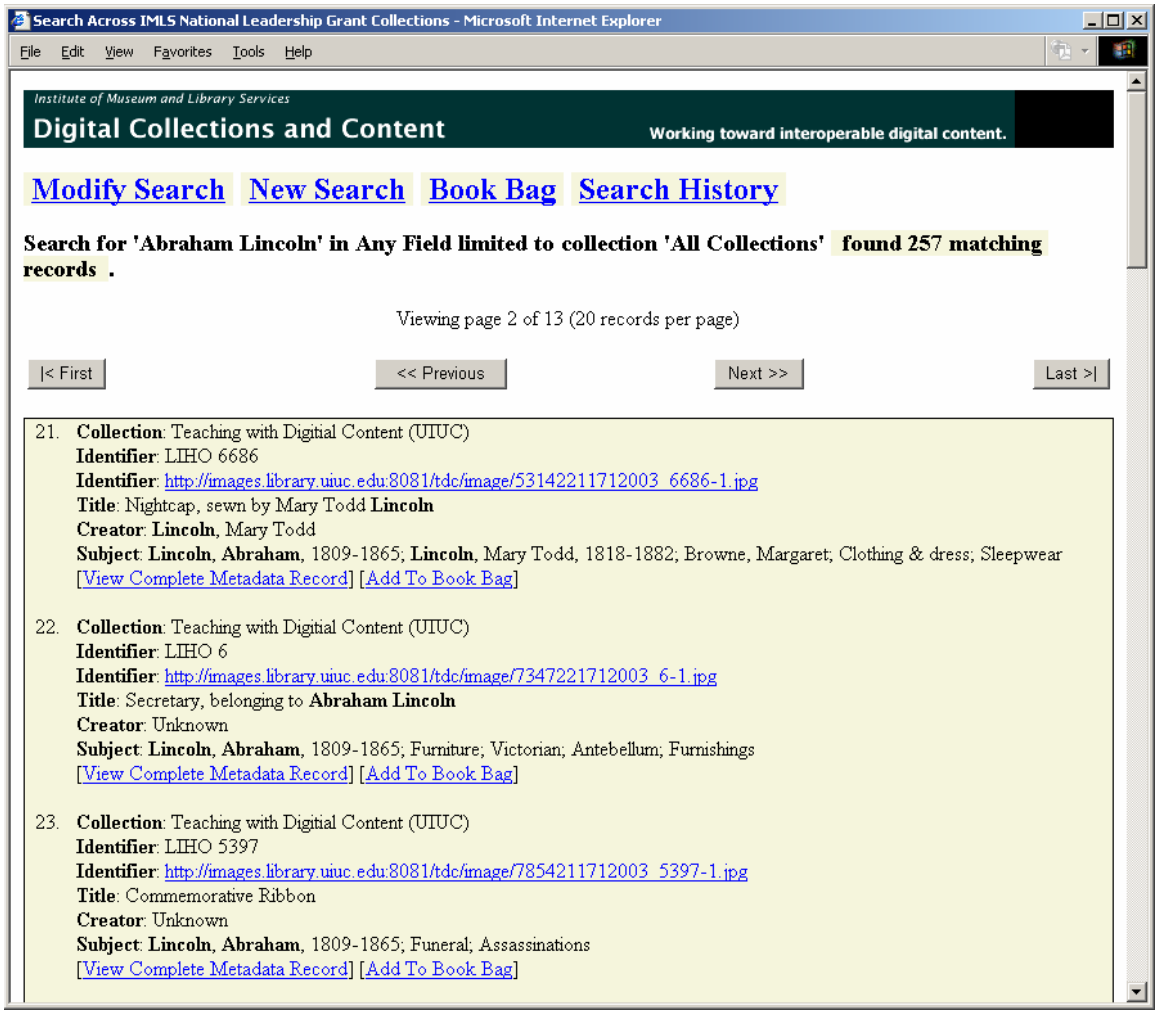

Figure 5 – Sample result list from simple search of IMLS DCC metadata repository (beta version).